RSAC 2025 Conference

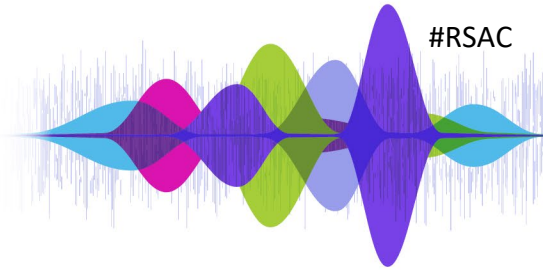
Many Voices.
One Community.

SESSION ID: PDP2-T09

Classify It: Developing a Data Classification Engine for Data Loss Prevention

Robin Franklin Guha

Security Engineer
Meta
https://www.linkedin.com/in/robin-franklin-b3388416a



Disclaimer

Presentations are intended for educational purposes only and do not replace independent professional judgment. Statements of fact and opinions expressed are those of the presenters individually and, unless expressly stated to the contrary, are not the opinion or position of RSA Conference LLC or any other co-sponsors. RSA Conference LLC does not endorse or approve, and assumes no responsibility for, the content, accuracy or completeness of the information presented.

Attendees should note that sessions may be audio- or video-recorded and may be published in various media, including print, audio and video formats without further notice. The presentation template and any media capture are subject to copyright protection.

© 2025 RSA Conference LLC or its affiliates. The RSAC and RSAC CONFERENCE logos and other trademarks are proprietary. All rights reserved.

Agenda

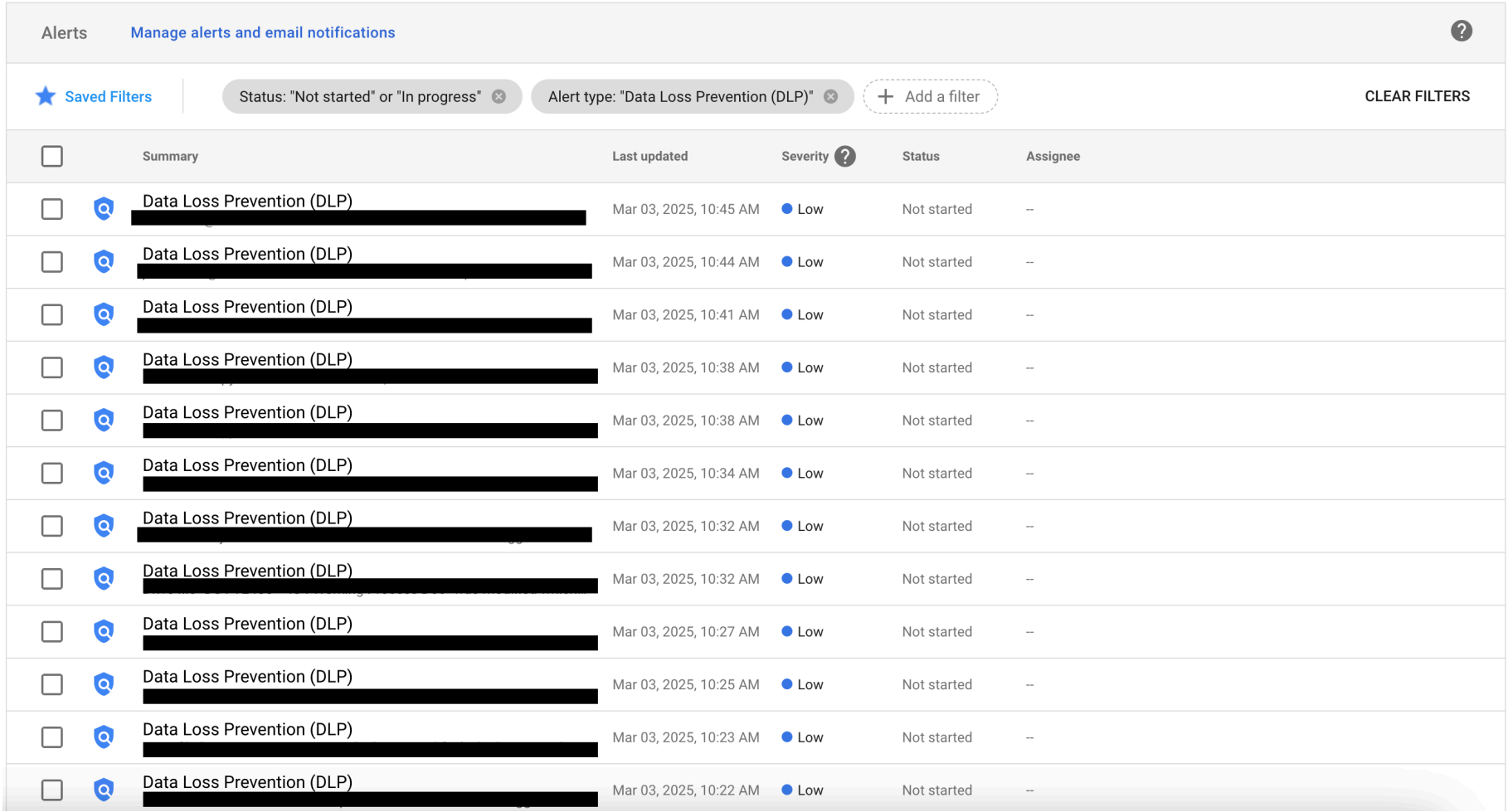
- Why Data Classification
- Data Sensitivity Levels
- Classification Approaches
  - Regex
  - LLM
  - Traditional ML
- Deploying and Integrating the Models

Data Classification

- Google and Microsoft Offer Data Loss Prevention Labels
- Reduced time for investigation and response -> reduced exfiltration

Source: Google Workspace Updates

Data Classification


Alerts
Manage alerts and email notifications
Saved Filters
Status: "Not started" or "In progress"
Alert type: "Data Loss Prevention (DLP)"
Add a filter
CLEAR FILTERS
Summary
Last updated
Severity
Status
Assignee
Data Loss Prevention (DLP) [redacted] Mar 03, 2025, 10:45 AM Low Not started --
Data Loss Prevention (DLP) [redacted] Mar 03, 2025, 10:44 AM Low Not started --
Data Loss Prevention (DLP) [redacted] Mar 03, 2025, 10:41 AM Low Not started --
Data Loss Prevention (DLP) [redacted] Mar 03, 2025, 10:38 AM Low Not started --
Data Loss Prevention (DLP) [redacted] Mar 03, 2025, 10:38 AM Low Not started --
Data Loss Prevention (DLP) [redacted] Mar 03, 2025, 10:34 AM Low Not started --
Data Loss Prevention (DLP) [redacted] Mar 03, 2025, 10:32 AM Low Not started --
Data Loss Prevention (DLP) [redacted] Mar 03, 2025, 10:32 AM Low Not started --
Data Loss Prevention (DLP) [redacted] Mar 03, 2025, 10:27 AM Low Not started --
Data Loss Prevention (DLP) [redacted] Mar 03, 2025, 10:25 AM Low Not started --
Data Loss Prevention (DLP) [redacted] Mar 03, 2025, 10:23 AM Low Not started --
Data Loss Prevention (DLP) [redacted] Mar 03, 2025, 10:22 AM Low Not started --

Data Sensitivity Levels

- Sensitivity Level 1: Public Information
  - Information already publicly available about the project or company
- Sensitivity Level 2: Internal Only
  - Information that should be accessible to everyone at the company
    - Ex: Benefits information
- Sensitivity Level 3: Confidential Information
  - Information that is highly sensitive and shouldn’t be open to everyone at the company
    - Ex: Intellectual Property, Security, Privacy, Legal documents
- Sensitivity Level 4: Personally Identifying Information
  - Information that personally identifies individuals
    - Ex: Employee performance reviews, candidate feedback

Data Classification

- What data sensitivity level would you give the following:

Document Title: Meta's Performance Philosophy

Purpose:

This document provides a summary of Meta's performance philosophy, outlining the company's approach to performance management and development.

Key Sections:

Introduction:
- Overview of Meta's commitment to performance excellence.
- Importance of aligning individual goals with company objectives.

Performance Management:
- Description of Meta's performance management process.
- Explanation of how performance is evaluated and feedback is provided.

Development Opportunities:
- Overview of training and development programs available to employees.
- Discussion of opportunities for growth and advancement within the company.

Core Values:
- Summary of Meta's core values and how they relate to performance.
- Explanation of how employees are expected to embody these values in their work.

Accountability:
- Discussion of the importance of accountability in performance management.
- Explanation of how employees are held accountable for their actions and results.

Data Classification

- What data sensitivity level would you give the following:

Document Title: Project Epsilon

Purpose:

This document outlines the framework for a strategic initiative, Project Epsilon, with a proposed launch date of October 2018

Key Sections:

Overview:

- Description of the project's objectives and scope.
- High-level timelines and milestones.

Partnerships:

- Discussion of potential partnerships and collaborations.
- Terms and conditions of these agreements.
- Proposed restructuring

Technical Requirements:

- Overview of the technical infrastructure needed to support the project.
- Security measures to protect sensitive data.

Financial Projections:

- Estimated costs and revenue projections for the project.
- Breakdown of expenses and funding sources.

Risks and Mitigation Strategies:

- Identification of potential risks and challenges.
- Plans for mitigating these risks and ensuring project success.

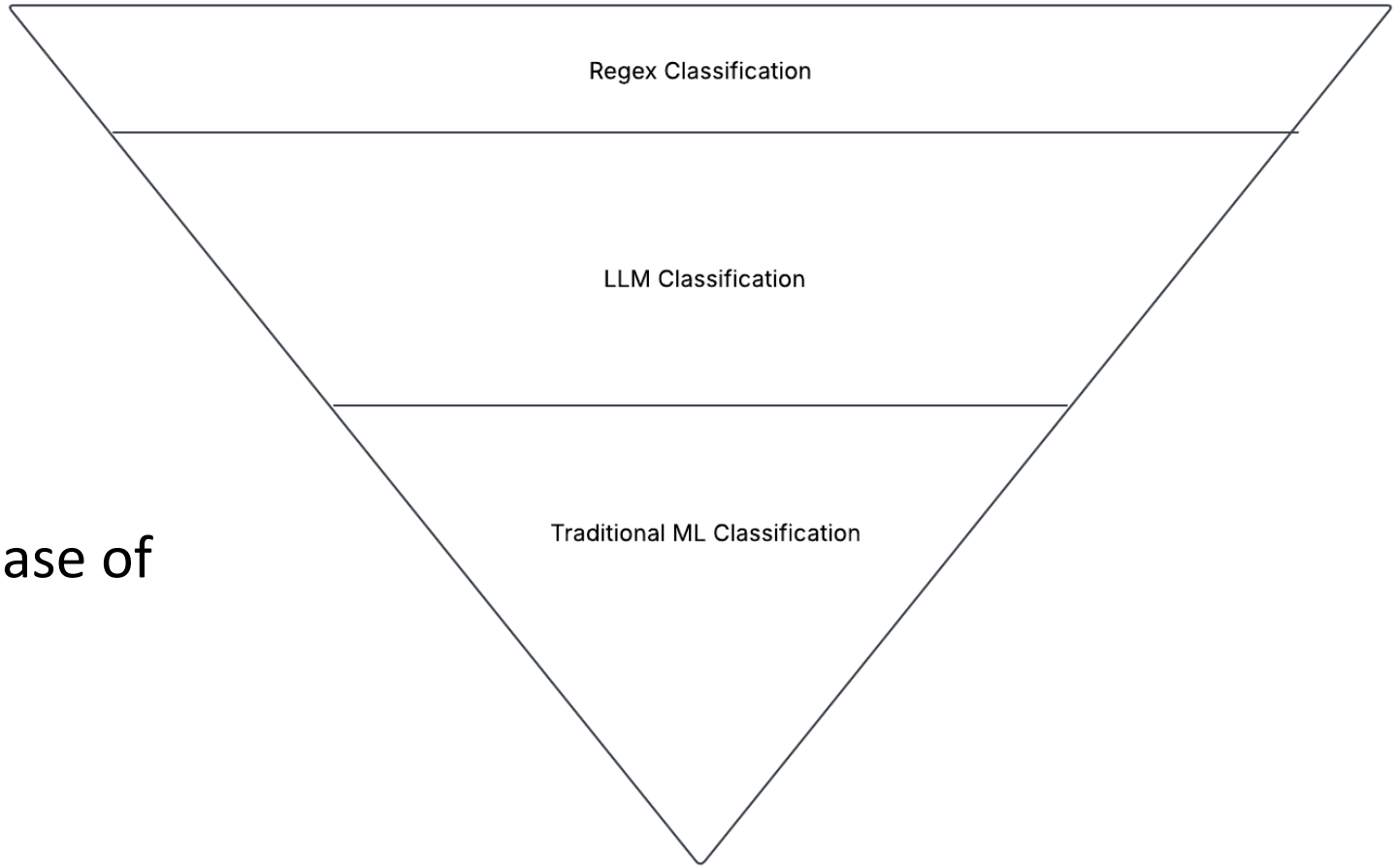
RSAC 2025 Conference

Classification Engine

Many Options, Different Effort Levels

Many Voices.
One Community.

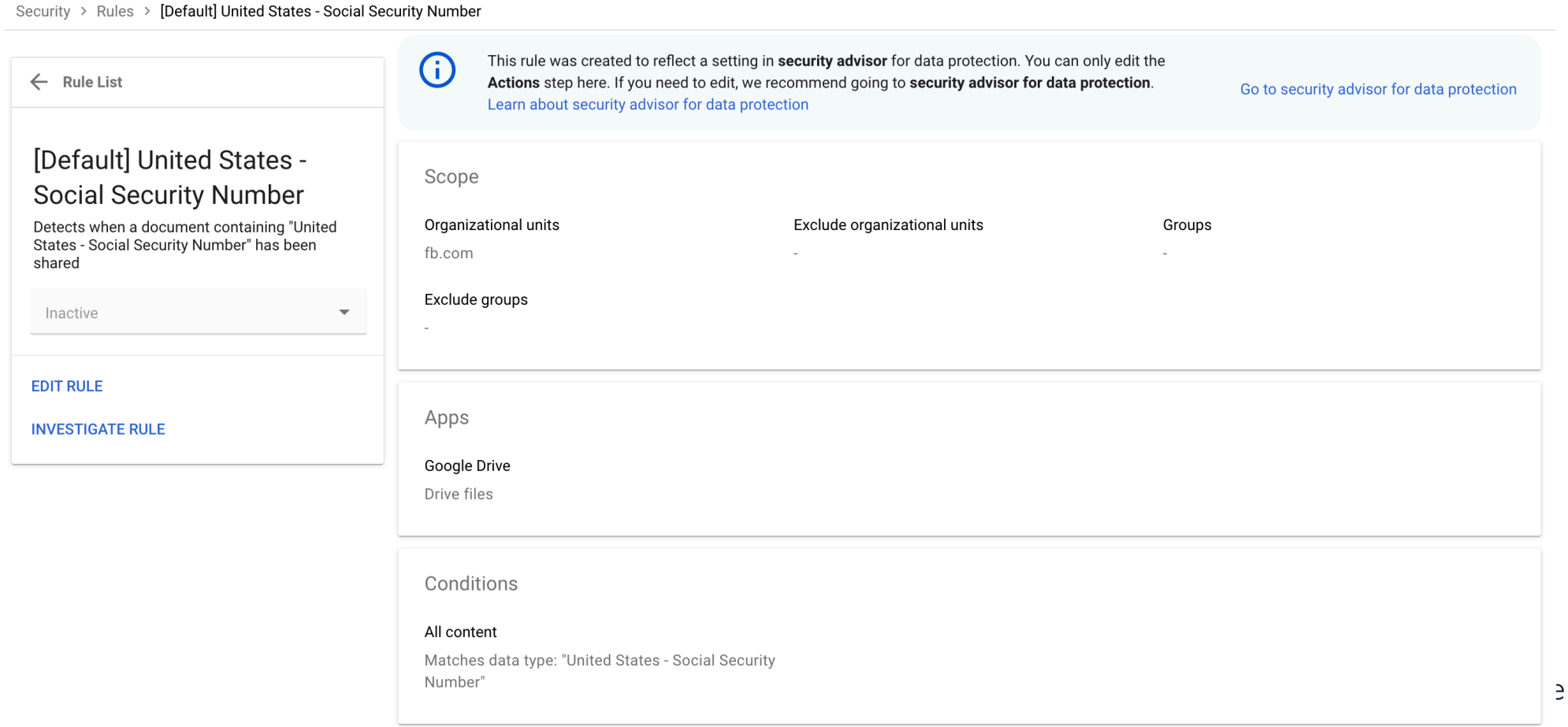
Data Classification Funnel



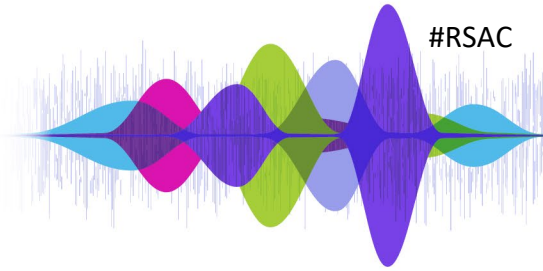
Note: In order of ease of implementation

V0: Regex

- Google and Microsoft DLP Engines built in Regex
  - Usually support file titles and content

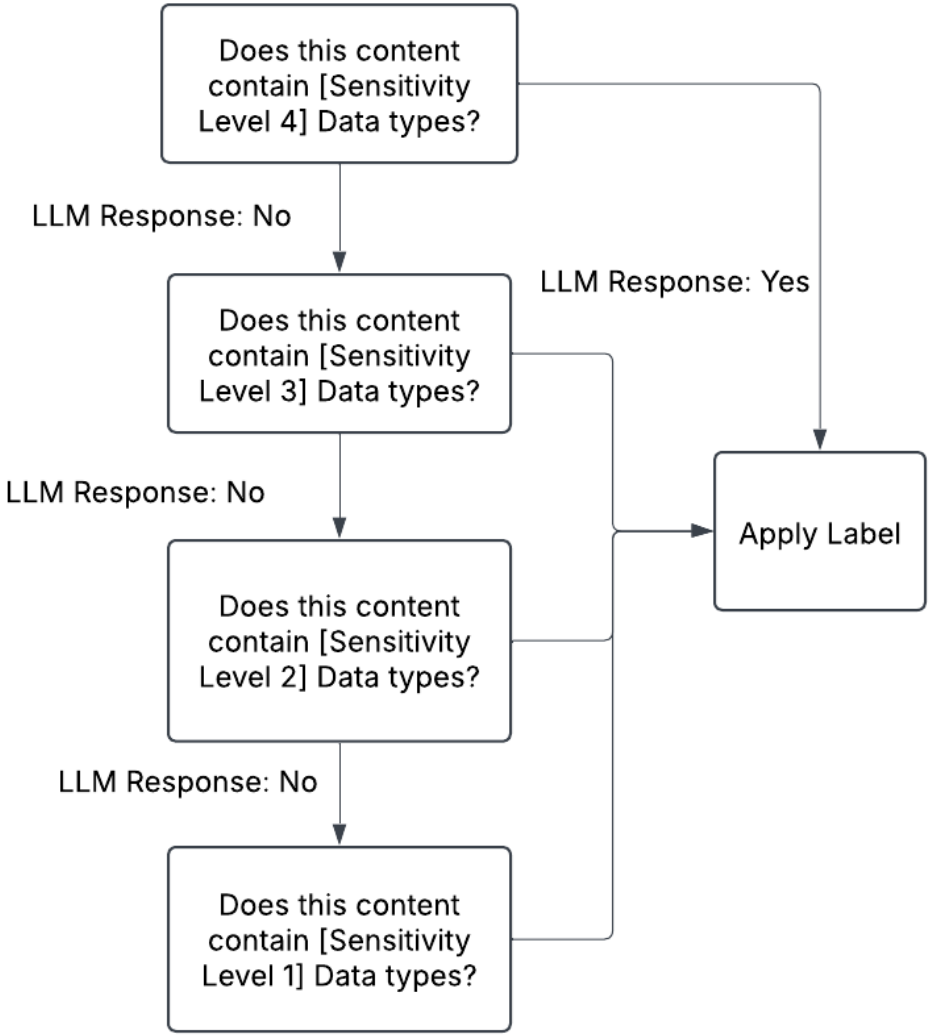

Security › Rules › [Default] United States - Social Security Number
Rule List
[Default] United States - Social Security Number
Detects when a document containing "United States - Social Security Number" has been shared
Inactive
EDIT RULE
INVESTIGATE RULE
This rule was created to reflect a setting in security advisor for data protection. You can only edit the Actions step here. If you need to edit, we recommend going to security advisor for data protection.
Learn about security advisor for data protection
Go to security advisor for data protection
Scope
Organizational units
fb.com
Exclude organizational units
-
Groups
-
Exclude groups
-
Apps
Google Drive
Drive files
Conditions
All content
Matches data type: "United States - Social Security Number"

V1: LLMs



- Vanilla LLMs deployed as an API Endpoint (e.g: AWS Bedrock, GCP Vertex AI)
- Keep It Simple – Decompose and Stack Prompts, Ordering of Prompts is Important to Minimize API Calls
- Requires an ingestion strategy
  - What events trigger a reclassification?
  - How would we approach back classification of files without labels?

V1: LLMs


Does this content contain [Sensitivity Level 4] Data types?
LLM Response: No
LLM Response: Yes
Does this content contain [Sensitivity Level 3] Data types?
LLM Response: No
Apply Label
Does this content contain [Sensitivity Level 2] Data types?
LLM Response: No
Does this content contain [Sensitivity Level 1] Data types?

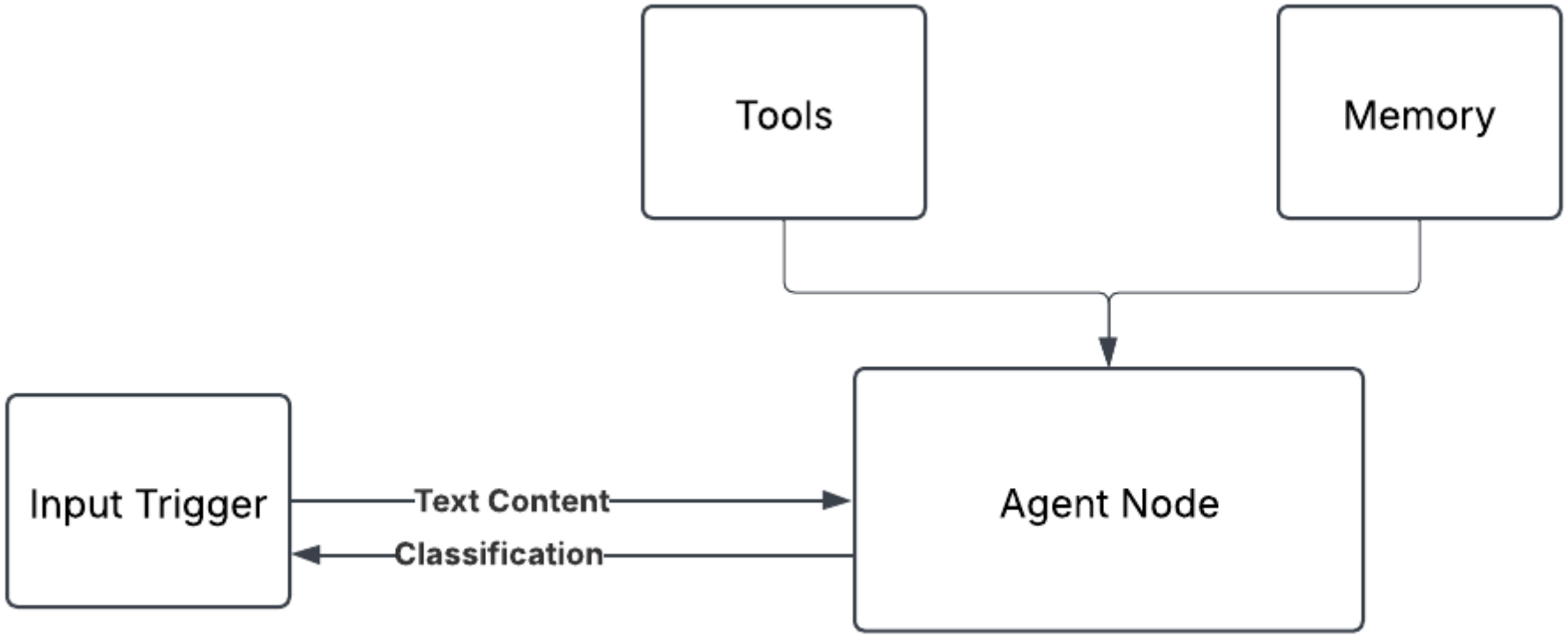
- Prompt Decomposition
  - Ask the LLM about data types, **don't** provide data sensitivity definitions

V1: LLMs

- System Prompt (Common for all Tasks)
  - You are an expert security analyst focused on determining the data sensitivity of text content. Carefully review the following:
- Task Prompt (per-sensitivity level)
- Sensitivity 4 Prompt:
  - Does this text contain any employee performance review information or compensation details? Does this text contain any details that would be found on a resume, such as an employer name and description of duties along with contact information?

V1.5 Agentic LLMs

- Like normal LLMs, but with more reasoning capabilities



V1.5 Agentic LLMs

- Key Tools:
  - Sensitivity Policy Lookup

```python
@tool
def policy_query_tool(
    self,
    query: str = "This is an example query",
    sensitivity_level: str = "Sensitivity level 1",
):
    """
    The policy_query_tool is a tool that queries data sensitivity level policy documents for details about a provided sensitivity level.
    """

    return policy_docs
```

V1.5 Agentic LLMs

- Generic Agent Node

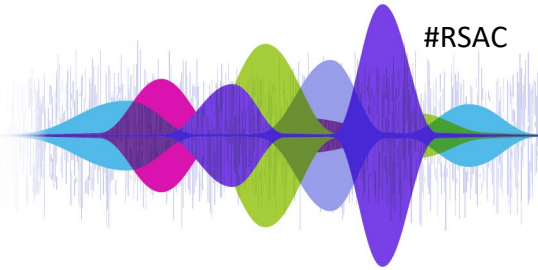
```
class ReActAgentNode(Node):
    def _prepare_prompt(self, prompt_template: str):
        if not prompt_template:
            prompt_template = """Answer the following questions as best you can. You have access to the following tools:

            {tools}

            Use the following format:

            Question: The input question you must answer or a task you must complete.
            Thought: You must always think about what to do next. You can think about the following actions.
            Action: the action to take, should be one of [{tool_names}]
            Action Input: the input to the action
            Observation: the result of the action
            ... (this Thought/Action/Action Input/Observation can repeat N times)
            Thought: I now know the final answer
            Final Answer: the final answer to the original input question

            Begin!

            Question: {input}
            Thought:{agent_scratchpad}"""
        return PromptTemplate.from_template(prompt_template)
```

V1.5 Agentic LLMs

- Classification Specific Task

```
task = f"You have been given the task to check if this document owned by SRC employee, background = {src_background}, contains sensitive information. Check thoroughly for different
react_node = ReActAgentNode(
    name="Sensitive Data Lookup Node",
    # Task Description
    task=task,
    tools=[policy_query_tool],
)
```

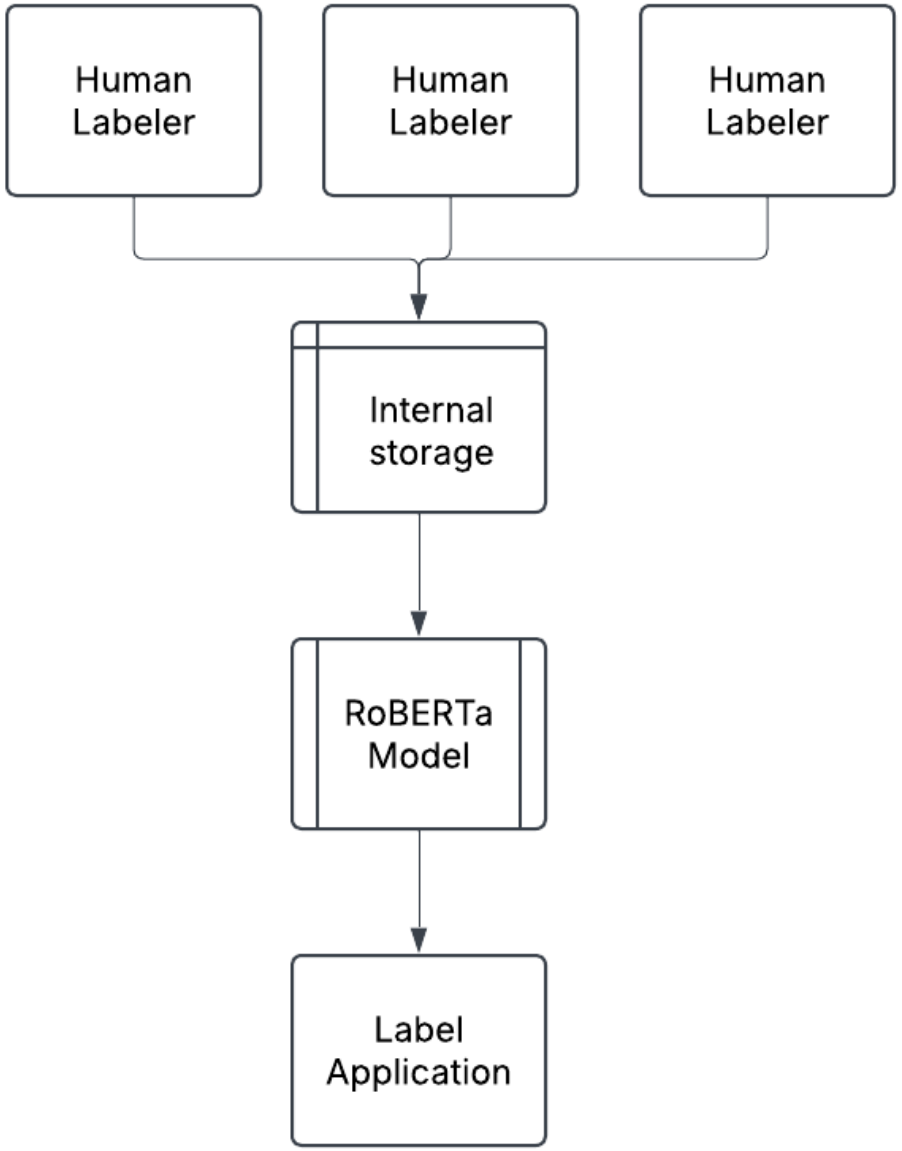
V2: Traditional ML



- End goal
  - Deterministic
  - Precise Precision and Recall Estimates at Thresholds
  - Requires precise training and test data
  - Least Infra needed (smaller models -> faster inference time, less GPUs)

V2: Traditional ML

- Human Labeling (with QA) is a **must**
- RoBERTa model -> BERT based, with more pre-training
- Optimization: Maximizing **Recall** at **Minimum Precision** Metric



Recap: Classification Options and Drawbacks

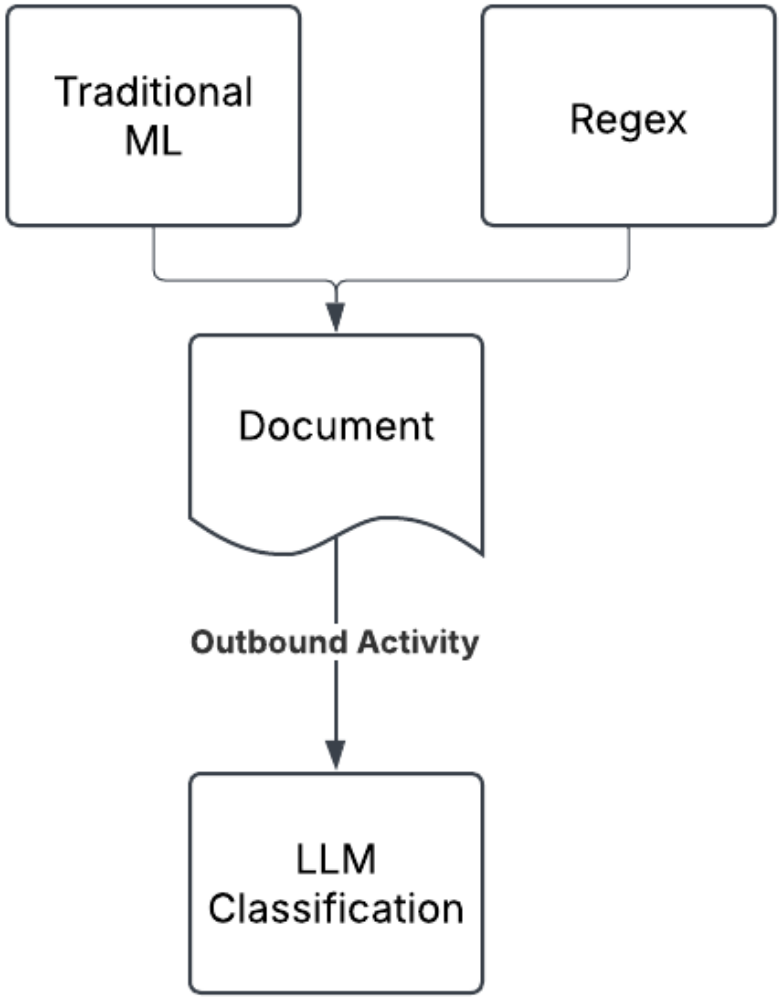
Classification Engine	Benefits	Drawbacks
Regex	- Easy, Inexpensive to Implement	- Noise

Classification Engine	Benefits	Drawbacks
Regex	- Easy, Inexpensive to Implement	- Noise
Vanilla LLMs	- Still pretty easy to implement and deploy - Higher Accuracy than regex	- Noise - Cost - Limited by number of questions and data types

Classification Engine	Benefits	Drawbacks
Regex	- Easy, Inexpensive to Implement	- Noise
Vanilla LLMs	- Still pretty easy to implement and deploy - Higher Accuracy than regex	- Noise - Cost - Limited by number of questions and data types
Agentic LLMs	- High Accuracy and Recall - Reasoning also surfaced	- Cost - Speed - Tuning of tasks and prompts

Classification Engine	Benefits	Drawbacks
Regex	- Easy, Inexpensive to Implement	- Noise
Vanilla LLMs	- Still pretty easy to implement and deploy - Higher Accuracy than regex	- Noise - Cost - Limited by number of questions and data types
Agentic LLMs	- High Accuracy and Recall - Reasoning also surfaced	- Cost - Speed - Tuning of tasks and prompts
Traditional ML	- Highest Accuracy - Deterministic - Well-defined Output - Speed	- Training Data Required - Model training time

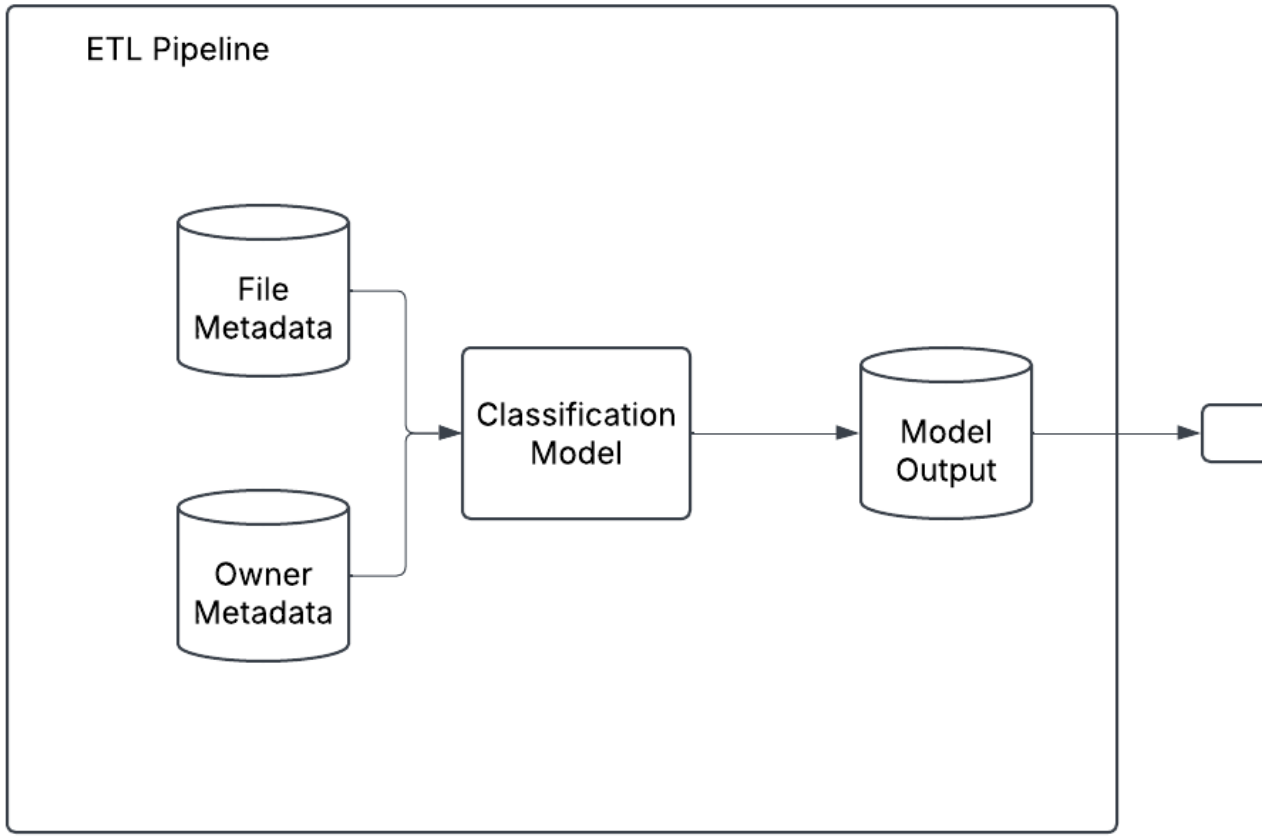
Ideal Data Classification Flow



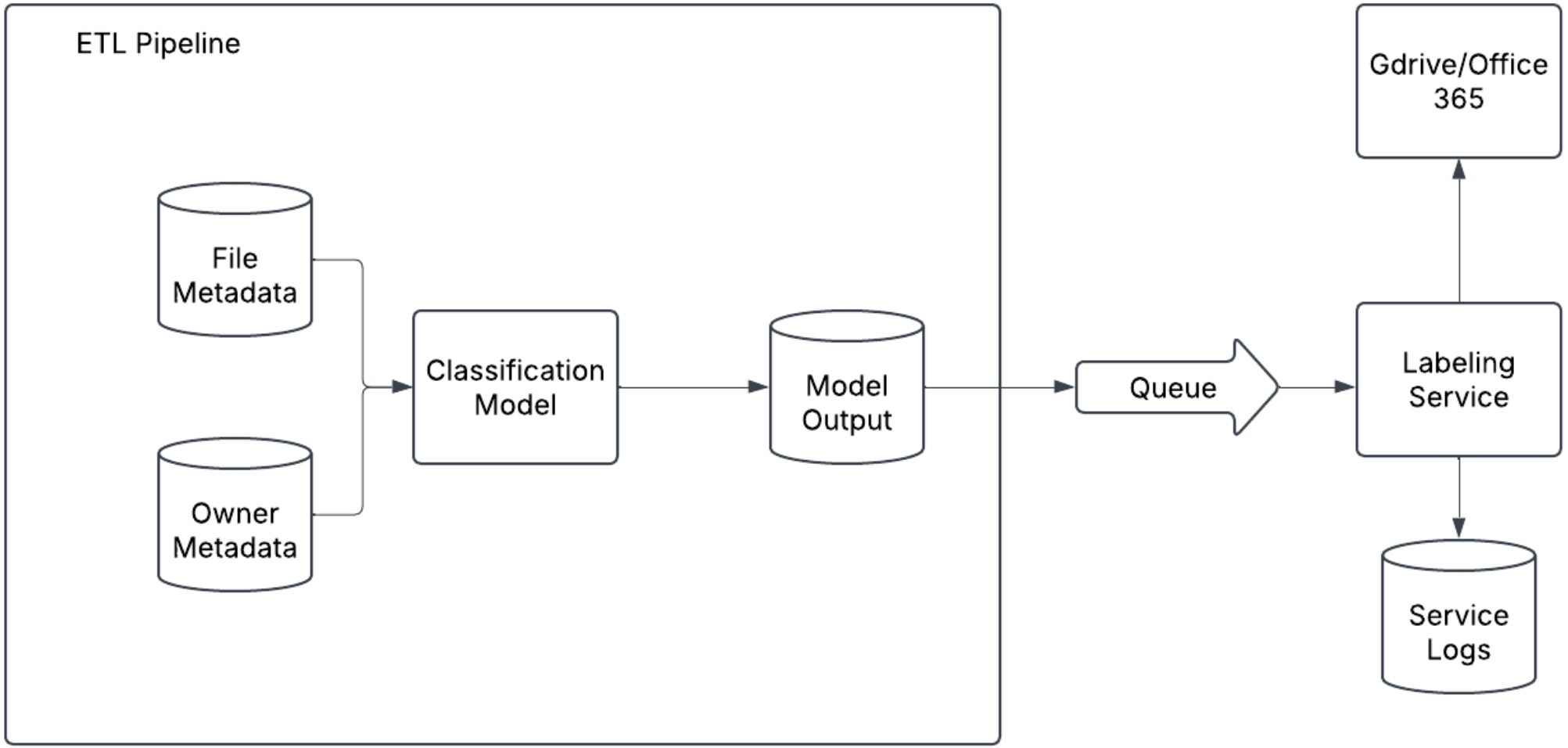
RSAC 2025 Conference

Classification Model Integration

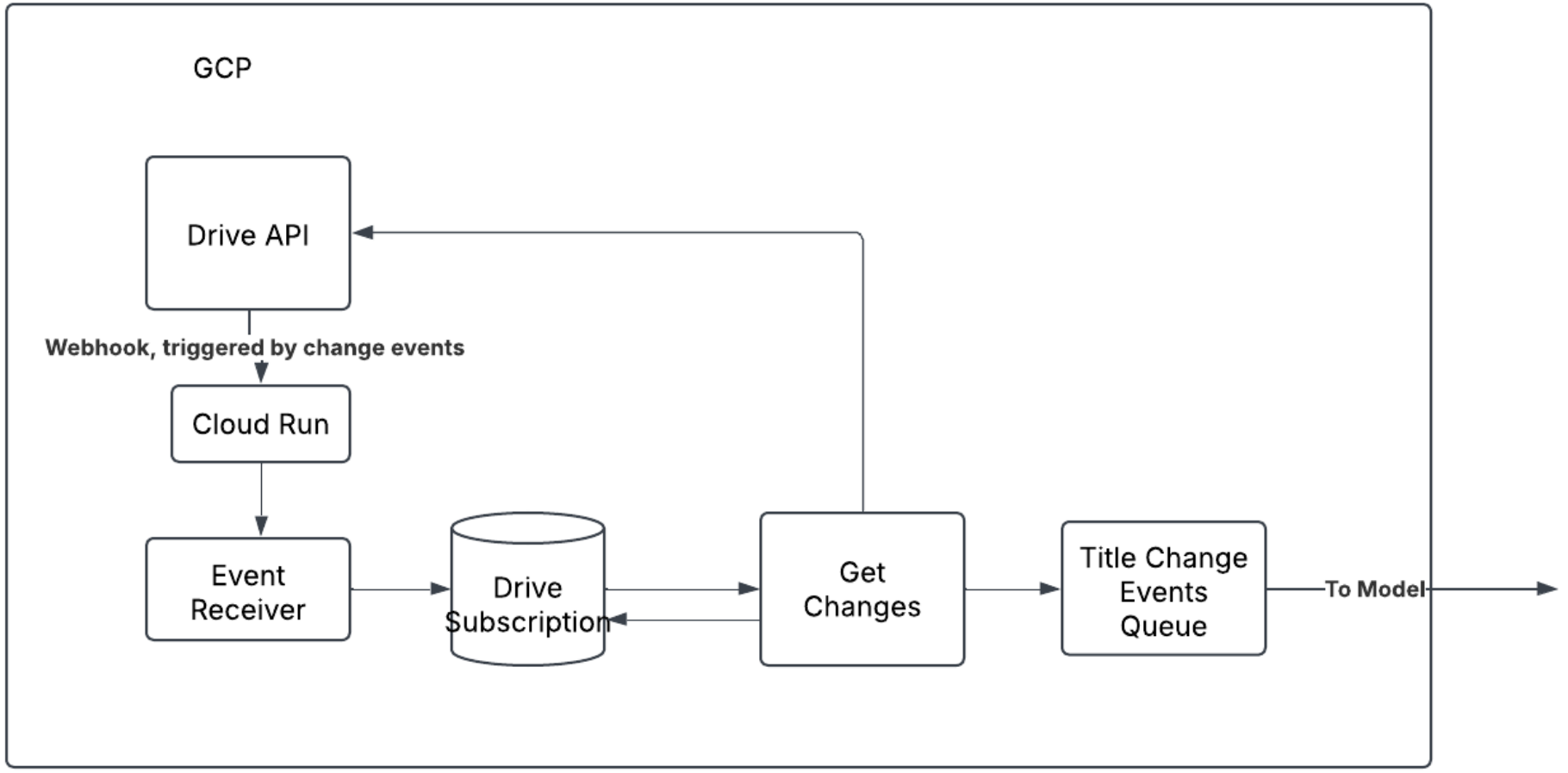
Batch Classification



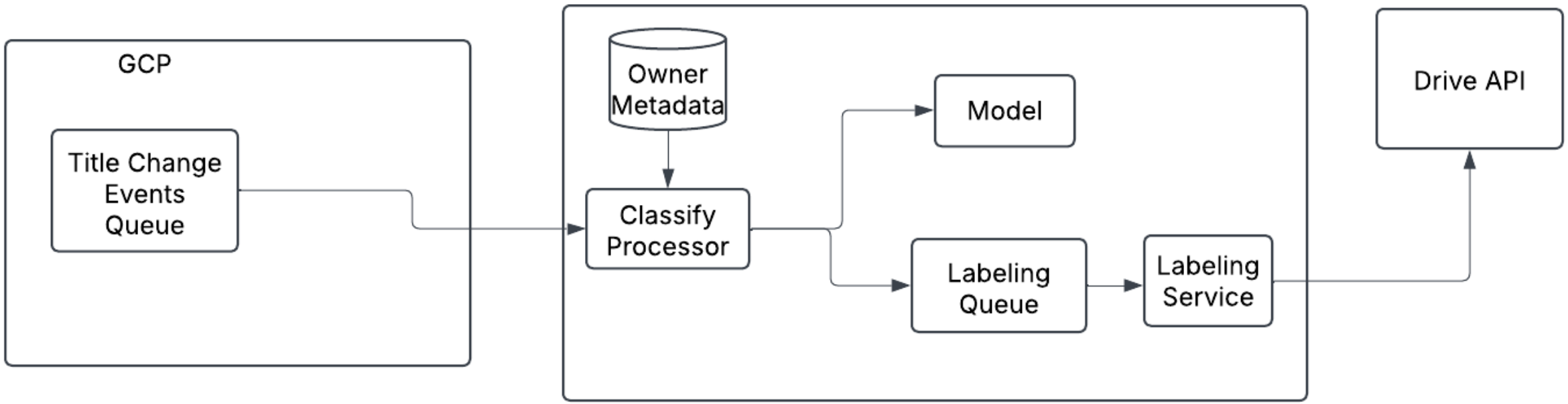
Batch Classification



Realtime Classification



Realtime Classification



Apply:

- Short Term
  - Identify applicable classification levels for your files
  - Begin implementing regex classifiers
- Medium Term
  - Begin iterating on LLM implementations, ID appropriate tasks and data types to include in prompts
- Long Term
  - Understand the quality of labeled data available to you
  - Begin iterating on traditional ML classifiers

Resources

- RoBERTa Overview
- Llama Stack - Deploying Llama as an Agent or a standalone Server
- Langchain – Agentic Examples

RSAC™ 2025 Conference

Many Voices.
One Community.

Special thanks to project partners: Brett Tegert, Anke Xu